TRENDS IN SEMICONDUCTOR MEMORIES

Yasunao Katayama

IBM Research, Tokyo

Despite their great market success, DRAMs have not kept pace with microprocessor improvements, so researchers are looking to advanced high-speed DRAM and merged DRAM/logic technologies to increase memory system performance.

A computer system's performance and power consumption are often determined not only by the characteristics of its data processing components but also by how well it transfers the right data to the right place. Thus, memory system design is becoming increasingly important. There are three major reasons for this. First, each new generation of CMOS technology provides a better speed/power trade-off, in parallel with transistor miniaturization. This lets us leverage the CMOS scaling to reduce the time and energy required for data processing. Nevertheless, many of the time and energy factors involved in data transfer are not reduced. For example, a typical CMOS gate delay in the "fan-out of four" condition, which was about 1 nanosecond 10 years ago, is now about 0.1 ns. Meanwhile, the time needed to transfer data 10 cm within a typical printed circuit board is independent of CMOS scaling and remains at about 1 ns.

Second, the use of parallelism can improve on-chip data processing performance—for example, through superscalar or VLIW (very long instruction word) architecture. However, interchip data transfer cannot exploit a large degree of parallelism because the number of chip-to-chip connections (the packaging) is limited.

Third, as memory density becomes greater, the fan-out and branching ratios required for the memory circuitry become larger, resulting in at least a logarithmic increase in the time required for address decoding and appropriate data path selection.

As a result, while microprocessor performance improves exponentially according to Moore's law, memory system performance lacks a corresponding improvement.

Memory performance is determined mostly by the choice of memory hierarchy (the use of cache memory and so on), the choice of memory bus architecture, and the performance of DRAM (dynamic random-access memories), the primary products used for main memory. Of course, many of the concepts used in present memory systems, and particularly the DRAM concept, did not exist early in the history of computing. Before the mid-1960s, computer memory systems consisted of cathode-ray storage tubes, ferrite cores, and thin magnetic films.[1,2] As semiconductor technology matured, semiconductor memories began replacing these preliminary devices. At first the standard memory cell implementation was a six-transistor SRAM (static random-access memory) cell, which is now used mostly for cache and battery-backup memory.

A breakthrough occurred with the invention of the one-transistor dynamic memory cell in 1968.[3,4] The idea combines a capacitor, for storing different amounts of charge to represent the distinguishable binary logic state, and a MOS transistor, for selecting a particular memory cell. A few years later, the DRAM became successful in computer main memory applications. Since then, thanks to the low bit cost and high density resulting from its simple cell design and the maturity in producing MOS VLSI (very large scale integration) chips,[5] the DRAM has dominated the computer main memory market.

DRAM success factors

Why have DRAMs remained the technology of choice for computer main memory for so long? The answer lies in their excellent architecture and sophisticated operation scheme, which have given them the highest density and the lowest bit cost among random-access memories. Figure 1 depicts a simplified memory architecture and the operation scheme for a typical FPM (first-page mode) DRAM design.

Cell design. The memory cell consists of only the components that are absolutely

0272-1732/97/$10.00 © 1997 IEEE

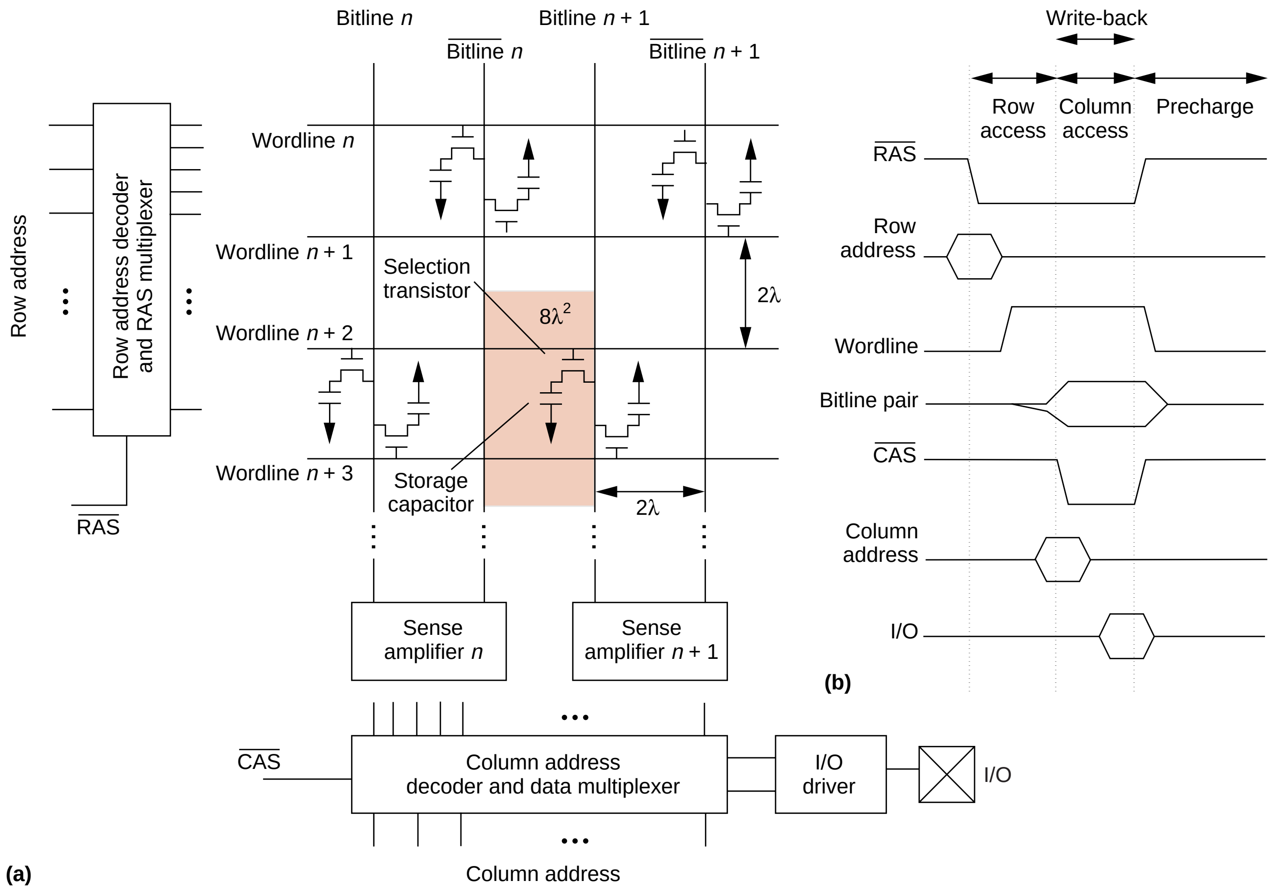

Figure 1. Memory architecture (a) and read operation scheme (b) for a typical DRAM with folded bitlines and conventional first-page-mode interface. The peripheral circuits are simplified.

necessary (storage capacitor and selection transistor). Its binary state is represented by the amount of charge it holds. Initially, the storage capacitor was implemented as a MOS planar capacitor. Even though the memory cell may look trivial, storing the data in a passive MOS capacitor was a revolutionary step, considering the level of MOS technology at the time (junction leakage and so on) and the absence of a proper sensing scheme for the passive capacitor. Nowadays, thanks to the simple cell structure, three-dimensional memory cell implementations using either trench or stack structure are widely used to maintain both memory cell scaling and a reasonable cell capacitance.

Array architecture. The simple cell architecture also allows the adoption of a cross-point array architecture for the memory array, realizing the high-density implementation of a solid-state RAM under lithography feature λ. The memory cell is defined by the intersection of two lines, one for the data selection (the wordline) and one for the data transfer (the bitline). In reality, the folded-bitline architecture, which is a modified version of the original cross-point architecture, is widely used. The folded-bitline architecture using bitline pairs takes at least $8\lambda^2$, as opposed to $4\lambda^2$ (assuming both bitlines and wordlines are placed in pitch 2λ) for the original cross-point architecture. Nevertheless, the common-mode capacitive noise during sensing is minimized by differential signal amplification using a balanced latch circuit. Since the memory cell and bitline pairs are massively parallel, a huge memory bandwidth (possibly more than 100 gigabits per second) is available in the array.

Operation scheme. The memory cell's read operation consists of row access, column access, write-back, and precharge.

- The row access starts the row address latch and decoding when the RAS (row address strobe) signal becomes low. This path can be considered as a multiplexing path of the RAS, which eventually activates an appropriate wordline according to the row address bits. The selected wordline connects the selected row of memory cells to the bitline pairs. The charge transferred from each memory cell to one of the corresponding bitline pair is amplified differentially and latched by the sense amplifier.
- The column access path is the combination of the column decoding and the multiplexing of the data latched in the sense amplifiers. A fixed number of bits in the sense amplifiers are selected and transferred to the

external bus according to the column address bits. The row and column address inputs are usually multiplexed to minimize the number of pins in the DRAM package.

- Since the read operation is destructive (that is, the cell itself cannot restore the original signal), the data in the selected row must be written back to the memory cells in parallel with the column access path. Even though the write-back doesn't affect the access time, it nevertheless causes a serious limitation in the RAS cycle time.
- The array must be precharged for the next memory access operation. Even though the column cycle can be purely static regarding sense amplifiers as SRAM cells, the dynamic memory cell requires proper precharge operations in the row cycle.

In the write operation, the bitline pairs are forced into a new logic state when the selected bitline pairs are connected to the external circuitry. The rest is basically the same as the read operation. In other words, writing to a DRAM cell is quite similar to writing back to the cell with a new logic state.

The ohmic charge transfer (the current is carried by the electron's microscopic diffusive process) in both reads and writes makes the operation principle independent of the circuit's dimensions. Therefore, the DRAM has been miniaturized in the macroscopic scale without altering the basic operation scheme. In addition, compared with flash or FRAM (ferroelectric RAM), the ohmic operation principle causes few significant material defects in consecutive read/write cycles.

DRAM limitations

Currently, the greatest limitation of DRAMs is their performance, and this involves two main aspects: the latency and cycle time in the row access and the data rate in the column access. The first issue is unique to DRAMs, while the second is a memory interface issue common to other types of semiconductor memory.

Performance in row access. In the random-access mode, where memory access starts with row access, performance is seriously constrained by the slow RC time constant in both charging and discharging the dynamic memory cell and array. Unlike an SRAM cell, where built-in positive feedback can restore the cell's own information, a DRAM cell is destructive, and the full amount of the signal must to be written back to the memory cell. The slow RC time constant results primarily from the cell's capacitance and the selection transistor's resistance. Capacitance is often kept higher than 10 femtofarads because of the soft-error problem. This problem occurs when the memory content is lost owing to the bombardment of the cell by alpha particles from cosmic rays and radioactive impurities in the package as well as in the chip itself. Resistance cannot be smaller than the value determined by the geometrical constraints in the width and length of the pitch-limited selection transistor in the memory cell.

The resistance problem becomes worse during write-back because of the asymmetry in the selection transistor operation. If the selection transistor is NMOS, it takes much longer to write back the high level than the low level. This is because the selection transistor is in source-follower mode for high-level write-back. (That is, the maximum level of cell voltage V_{cell}, which acts as a source of the NMOS transistor, follows the wordline voltage minus threshold voltage $V_{WL} - V_T$.) Therefore, the selection transistor's resistance is much higher because of a smaller gate overdrive. Even though wordline boosting or limited bitline swing (using a higher V_{WL} than V_{BL}) helps reduce the asymmetry, this is the most fundamental performance limitation in the DRAM random-access cycle.

In addition to the RC constant in the memory cell, the RC time constant in the array, which results from the driver hierarchy's large fan-out and branching ratios, further compounds the performance problem. Although this issue is not unique to DRAMs,[5] the situation is worse than in SRAMs because of the DRAM's higher density and the dynamic nature of the DRAM cell. The memory cell is arranged in one dimension and lacks internal structure from the logical point of view. Nevertheless, in the physical implementation the enormous fan-out and branching ratios are handled by distributing them in the combination of the row and column using the two-dimensional nature of the cross-point architecture. Still, even though from the performance aspect it is optimal to maintain a ratio of 3 to 4, the area constraint requires the use of larger ratios, particularly within the memory array. To reduce the number of array circuits, such as wordline drivers and sense amplifiers, designers often connect 128 to 1,024 selection transistors in each wordline and 128 to 256 cells in each bitline pair (excluding the redundancy). For example, the signal out of memory cell V_{sig} is reduced by charge-sharing when the memory cell is connected to the bitline. V_{sig} is given by

$$V_{sig} = V_{pre} + (V_{cell} - V_{pre}) * C_{cell} / (C_{cell} + nC_{BL}),$$

where V_{sig} is the voltage for the sensing signal, V_{pre} is the bitline precharge, V_{cell} is the cell storage signal, C_{cell} is the cell capacitance, C_{BL} is the bitline capacitance per cell, and n is the branching ratio. V_{sig} decreases as n increases, and thus the sensing takes longer. To handle the larger fan-outs and branching ratios of today's high-density DRAMs, designers commonly use more levels of hierarchy (for example, a subarray) in the design.

Moreover, the peak current during sensing also limits performance. This current results from concurrent charging and discharging of the huge capacitance due to the massively parallel architecture. Faster array operation requires a greater peak current and thus results in both greater power consumption and a larger voltage drop owing to the resistance and inductive effects in the power-supply paths.

Performance in column access. The cycle time in column access determines the data rate once the data have been latched in the sense amplifiers. Here, we need data multiplexing (in case of write) or demultiplexing (in case of read) according to the (partially) decoded column address bits. Appropriate width and frequency conversion between the sense amplifier latches and external I/Os are also needed. Widening the chip-to-chip connections increases the cost by increasing chip area and package size, because the number of I/O drivers and packaging pins increases. The simultaneous-switching-noise problem also requires a larger number of supply connections. On the other hand, increasing the bus frequency in the chip-to-chip interface requires the adoption of low-impedance terminated bus lines with reduced voltage swing. Even though the reduced voltage swing alleviates the power increase caused by the increased data rate, the increase in standby power resulting from the bus termination may limit the use of higher frequencies in certain applications without appropriate power management.

Memory refresh. Another limitation is the memory refresh problem, which is unique to DRAMs. Since the memory cell is dynamic, junction leakage and subthreshold leakage in the cell capacitor and selection transistor require that the cell's contents be refreshed at certain intervals (typically, every 16 to 32 milliseconds). Even so, the cell's dynamic nature is not a serious problem in present computer systems, thanks to the use of the magnetic hard disk drive as a permanent storage device. However, in other applications, building a nonvolatile or even a battery-backup memory system is difficult unless an SRAM or flash memory is used.

From technology push to application pull

Now that we've looked at a DRAM's advantages and limitations, let's review DRAM architecture evolution to see how performance limitations have been overcome. Figure 2 summarizes the major DRAM components in terms of off-chip data rate and functionality. Conventional DRAM development—mostly a density increase driven by the technology itself—is starting to aim more for a wide variety of products driven by application needs. Products along the data rate axis are general-purpose high-speed DRAMs, which boost performance by improving the off-chip data rate. Products along the functionality axis result from developing application-specific, high-functionality DRAMs. These improve performance by exploiting the DRAM's massively parallel data architecture (that is, by placing logic circuits before completely demultiplexing the data bus).

High-speed-DRAM development

Research on high-speed DRAMs started in the late 1980s at IBM[6] and Hitachi.[7] The resulting chips demonstrated a random-access time in the low 20-ns range and a column access time in the low 10-ns range, which made them two to three times faster than the 1-megabit DRAMs existing at that time. IBM's DRAMs realized the high-speed memory access through improved address latch and decoding circuits, memory array architecture, a sensing scheme, and data path circuits using CMOS technology. In particular, designers reduced the RC time constants in the array by using the second-level-metal strap for the wordline and boosting the wordline during writeback. They minimized operational power and noise using the half-V_{DD} sensing scheme. Hitachi's DRAMs realized high-speed memory access by combining the large-gain bipolar transistor and CMOS circuits using the Bi-CMOS process.

At that time, microprocessors were not fast enough to require these high-speed memory products; a relatively small amount of cache memory could satisfy the memory bandwidth requirement. Thus the first-generation high-speed DRAMs were successful as a research milestone but not in commercial terms.

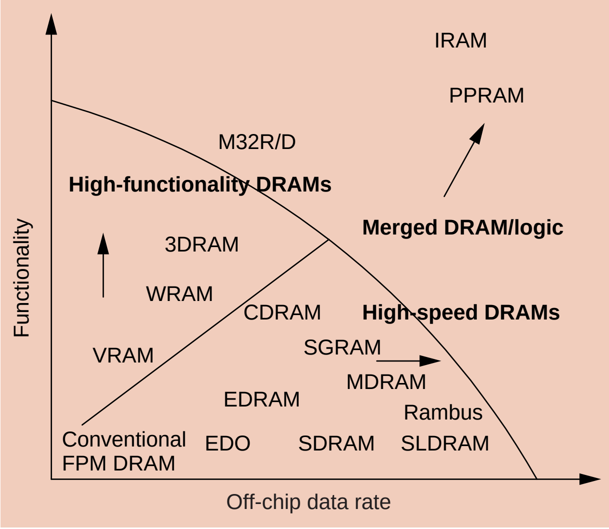

Figure 2. The industry's various DRAMs fall into three categories: high-speed DRAM, high-functionality DRAM, and merged DRAM/logic.

Faster data rate. In the 4-Mbit generation, EDO (extended data out) DRAMs improved the data rate for column access by adding an extra pipeline stage in the output buffer while keeping the conventional asynchronous interface. The difference is that the data are still valid at the end of the present CAS (column access strobe) signal and remain valid until the next column cycle starts. System designers liked this memory because it gave them improved performance with only minor modifications to the memory controller chipset. A typical 4-Mbit EDO DRAM with ×8 I/Os operating at a 33-MHz column access cycle can provide a peak data rate of 266 Mbytes/second per chip.

In the 16-Mbit generation, SDRAM (synchronous DRAM) employed a high-speed synchronous interface. By either pipelining the data path or prefetching several bits at a time over wider data lines, designers improved the data rate for column accesses. They also improved random-access performance by interleaving multiple banks of memory arrays (normally two) on the same chip. The peak data rate for a 16-Mbit SDRAM with ×16 I/O operating at 66 MHz is 1.1 Gbps (133 Mbytes/s) per chip. The burst mode avoids the provision of column addresses for every data transfer, even though changing the burst length requires an extra cycle.

JEDEC (the Joint Electron Device Engineering Council) has standardized both EDO DRAM and SDRAM.

Faster random access. Several development efforts focused on improving DRAM row access performance to near SRAM level by integrating a small amount of SRAM or by dividing the DRAM into multiple independent banks. EDRAM (enhanced DRAM)[8] has a distributed cache, while the cache of CDRAM (cached DRAM)[9] is rather localized. Although the standard DRAM can use sense amplifiers as a distributed cache, the extra buffers in EDRAM and CDRAM can improve the hit ratio. Both are available as product chips. In MDRAM (multibank DRAM),[10] row access is improved by integrating many independent 256-Kbit DRAM banks on a single chip for graphics and cache-memory applications. The chip realizes an average random-access time that approaches column access time by independently operating individual banks and thereby overlapping or hiding the row access and precharge.

Revolutionary memory interface. Rambus DRAM[11] uses a packet-type memory interface. The initial version realizes a peak data transfer rate of 500 Mbps per data I/O pin (and 4 Gbps, or 500 Mbytes, per chip) with a 250-MHz clock by transferring packets at both clock edges. The high data rate per chip is suitable for applications such as game machines and graphics cards, where the memory granularity requirement is small but the data bandwidth requirement is high. In standby, the DRAM array is active (the data are latched in the sense amplifiers), and array precharging occurs only when the new row address differs from the previous one, that is, in page miss.

The improved version, called Concurrent Rambus, realizes a peak data rate of 600 Mbps per data I/O pin with a more efficient protocol to achieve a higher effective data rate. Address and data multiplexing reduces the number of I/Os, even though the maximum data rate available for the data transfer decreases, particularly for scattered accesses (consecutive short-burst accesses in random addresses).

High-functionality-DRAM development

Instead of making the DRAM faster, designers responded to the needs of special applications—particularly graphics—and followed a different trend in DRAM development.

Video applications. The first commercially successful high-functionality DRAM product was VRAM (video RAM). Introduced in the mid-1980s, it realized concurrent dual-port access by multiplexing the massively parallel data in the DRAM array before the data were demultiplexed in the column data path. The video screen refresh data stored in the sense amplifiers are first transferred to serial registers located near the sense amplifiers and are then transferred to the screen via a separate serial-access port, which is different from the ordinary random-access port. Therefore, there is little interference between the random and serial accesses. With the 4,096-bit full transfer between the sense amplifiers and the serial registers, operating at 10 MHz, the internal data rate is 41 Gbps (5.1 Gbytes/s), a level not achieved in other high-speed DRAM products. However, VRAM is expensive, since placing serial registers near the sense amplifiers complicates the design and consumes die area. VRAM is JEDEC standard.

More functions. WRAM (window RAM)[12] improves graphics operations in a GUI environment. Instead of distributing registers near the sense amplifiers, as in VRAM, designers placed localized registers for serial access outside the DRAM array. Data for screen refresh travels over the 2.1-Gbytes/s internal bus. Therefore, at the cost of increased serial access overhead, the design became smaller and simpler, resulting in a reduced cost relative to VRAM. WRAM has additional functions such as aligned bitblt and block write with bit and byte masking.

The 3DRAM[13] was specially designed to accelerate the back-end image layer processing of 3D graphics applications. It changes the read-modify-write operation, which occurs frequently in 3D graphics applications, into a single write operation by integrating a pixel ALU for Z-compare and alpha-blending functions. Registers for serial access are located outside the array, as in WRAM.

Merged DRAM/logic technology

In terms of computer architecture, the integration of logic and memory goes back to Stone's research in 1970.[14] However, research on integrating DRAM and ASIC (application-specific integrated circuit) logic is relatively new.[15] Nevertheless, products with merged DRAM/logic technology are already found in the integrated graphic controller plus frame buffer chip used in many notebook computers and in the integrated CPU plus DRAM chip (the Mitsubishi 32R/D)[16] targeted at embedded applications. Moreover, DRAM and logic integration has already been achieved in high-functionality DRAM design. It is possible to take an extreme view and say that merged DRAM/logic technology has already been achieved in standard DRAMs, since they also contain logic circuits such as address decoders and latches. Nevertheless, further efforts are needed to encourage wider use of this technology.

Design methodology. It is important to develop a DRAM macro that is competitive in terms of flexibility and density and also make the macro usable in a standard ASIC logic design environment. A DRAM macro is a combination of a DRAM array and the necessary support circuits to form a functional memory unit with smaller memory granularity for integration with other logic circuits. The interface and the number of I/Os can be very different from those of stand-alone DRAMs because a DRAM macro is free of packaging constraints. The design challenge is to handle the trade-off between realizing configuration flexibility in, for example, the numbers of rows and columns and I/O width, and achieving a density comparable to that of stand-alone DRAMs.

VLSI architecture and design. The key design challenge is to create a chip architecture that utilizes the bandwidth available from the massively parallel data architecture in the array. Even though the raw bandwidth in the array is huge, the on-chip data bus architecture must be efficiently arranged for improved flexibility in connecting DRAM arrays and logic circuits. One method is to use a cross-bar architecture.

System design. Merged DRAM/logic technology has the potential to advance computer system architecture in ultra-high-density DRAMs beyond the 256-Mbit and 1-Gbit generations. University research on IRAM[17,18] (intelligent RAM) and PPRAM[19] (parallel processing RAM) is exploring such architectures. IRAM combines processors and DRAM technology to provide high bandwidth for predictable access, such as matrix multiplication, and low latency for unpredictable access, such as database access. (This assumes the use of 1-Gbit-level merged DRAM/logic technology.) PPRAM integrates four processing elements, each consisting of four 32-bit RISC processors, an 8-Mbyte DRAM, and a 24-Kbyte cache SRAM, using a 256-Mbit-level merged DRAM/logic technology.

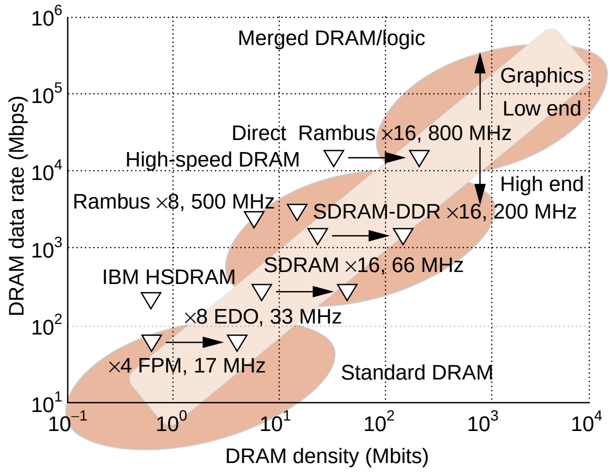

Figure 3. The relation between a DRAM's density and its peak data rate. The shaded band shows the area where major applications exist. Horizontal arrows indicate that the data rate becomes relatively low when the density quadruples.

What's driving DRAM changes?

Typically, the change in DRAMs is attributed to their ever-increasing performance gap relative to microprocessors. But there's another possible explanation. We start from the following relations, comparable to the Case/Amdahl rule-of-thumb:

$$\text{MPU performance} = k_1 \text{ (memory system bandwidth)}$$
$$= k_1 \text{ (DRAM data rate) } N_{\text{DRAM}}/N_{\text{bank}},$$

and

$$\text{MPU performance} = k_2 \text{ (memory system capacity)}$$
$$= k_2 \text{ (DRAM density) } N_{\text{DRAM}},$$

where N_{DRAM} is the number of DRAMs in the system, N_{bank} is the number of banks sharing the same data bus, and k_1 and k_2 are coefficients. By dividing the two equations, we get

$$\text{DRAM data rate} = N_{\text{bank}} \; (k_2/k_1) \text{ (DRAM density).}$$

This equation implies that the DRAM needs a higher data rate as its density increases. For example, if designers replace four 4-Mbit DRAMs with one 16-Mbit DRAM, the 16-Mbit DRAM should have a data rate four times higher to keep the same memory system performance. The coefficient N_{bank} (k_2/k_1), which is identical to what is called *fill frequency*,[20] depends on the application. In the current design, the empirical numbers are 100 Hz for PC graphics, 10 to 20 Hz for PC main memory, and less than 10 Hz for servers and workstations. The difference in coefficient results from larger k_1 and smaller k_2 in larger computer systems, owing to more powerful cache memory hierarchies and the relatively larger memory system capacity. The latter result from a larger working data set and larger applications.

Figure 3 shows the peak data rate and density of actual DRAMs, as well as a region where applications exist, assuming constant $N_{\text{bank}}(k_2/k_1)$. The drivers of DRAM changes are low-end and graphics applications, where the data rate requirement per DRAM density is higher. In fact, the major transition to high-speed DRAMs occurred not in the high end of the market but in the low end. The high-speed DRAM could provide smaller memory granularity for a given bandwidth requirement.

Where are DRAMs headed?

If we extrapolate the trend, the higher data rate requirement as DRAM density increases will necessitate a new generation of high-speed DRAMs. This requirement is expected to eventually drive a transition from high-speed DRAMs to merged DRAM/logic chips in many applications.

There are three major candidates for the next generation of high-speed DRAMs. The first—a near-term solution—is likely to be the SDRAM-DDR (SDRAM double data rate), which uses a synchronous RAS and CAS interface comparable to that of the original SDRAM. The data rate will be improved by transferring data at both edges of the clock. A 16-Mbit SDRAM-DDR with ×16 I/O operating at a 100-MHz clock (200-MHz data rate) can provide 3.2 Gbps.

The second candidate, Direct Rambus DRAM, is a further improved version of the Rambus DRAM line and is expected to provide a peak data rate of 13 Gbps per chip. The data rate is due to the 400-MHz clock (800-MHz data rate) and 16-bit bus width.

The SLDRAM, the third candidate, originates from Ramlink, IEEE standard 1596.4, which was developed by applying the SCI (scalable coherent interface) to the memory bottleneck problem. Originally, it had several new features, such as a unidirectional interface and point-to-point connections. There were two links: one for incoming signals of the command, addresses, and write data and the other for outgoing read data signals. However, the revised version[21] uses a regular bidirectional bus for the data, because often there are long bursts of write or read data that can disturb the balance between incoming and outgoing links.

Beyond high-speed DRAMs lies merged DRAM/logic technology. This transition will also be driven by power consumption and the need for a small footprint, particularly for mobile applications. Power consumption in the memory system is increasingly attributable to both intra- and interchip data transfer. It is becoming very important to reduce this power consumption without sacrificing memory bandwidth between the microprocessor and the DRAM.

The transition to the merged DRAM/logic technology is accelerated by smaller memory system capacity in low-end applications. As Figure 4 (next page) shows, if the memory density of each DRAM chip is larger than the memory system capacity, there are few alternatives to using the merged DRAM/logic technology. This situation will probably become more common, since the memory system capacity increases more slowly than the DRAM density, assuming that the cost for a certain kind of system will not change and that the bit cost will not decrease as rapidly as the DRAM density increases.[22] In other words, the chip costs more for each DRAM generation, so unless the number of DRAMs decreas-

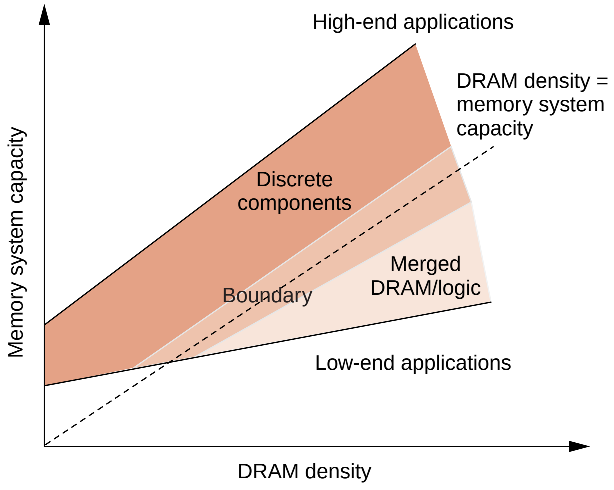

Figure 4. Relation between DRAM density and memory system capacity. Once the memory system capacity becomes smaller than the DRAM density, systems using merged DRAM/logic technology become preferable to those made of discrete components.

es, it is difficult for total system cost to remain unchanged. Of course, fluctuations due to the silicon cycle will cause cost perturbations, but this assumption seems reasonable from a long-term viewpoint.

Nevertheless, the transition may slow down, for both technical and commercial reasons. From a technical point of view, since DRAM and logic semiconductor technologies are different, merging the two at a reasonable cost still presents many challenges. The progress of system- and software-level solutions, such as advances in compilers, may increase cache memory efficiency and thus reduce the DRAM data rate somewhat (as shown in the more lightly screened area in Figure 3). Progress in packaging technology may also delay the transition. From a commercial point of view, memory users care about the technical standard and second source for a secure, stable, high-volume supply. Therefore, the transition will not occur until the merged DRAM/logic technology can provide an above-threshold improvement over alternative, incremental solutions.

The increased level of integration also drives the transition from high-functionality DRAMs to merged DRAM/logic technology. As DRAM density increases, it becomes possible to integrate more logic circuits onto the same chip. Thus, the limitation in designing complex custom logic circuits will drive a transition to the merged DRAM/logic technology.

Assuming that the limitation in merging logic and DRAM will not disappear soon, one good approach will be the DRAM-based coprocessor to achieve a better trade-off between the architecture and technology issues. This approach, shown in Figure 5, involves a DRAM-based coprocessor containing on-chip massively parallel logic accelerators for *data-intensive* processing. Here, the performance is limited not by the data processing speed but by the memory bandwidth, and parallel implementation of the processing units with the DRAM will greatly improve total throughput. The design uses a DRAM-based merged DRAM/logic technology, while a separate microprocessor (using high-speed logic technology) helps to achieve the best logic speed for *computation-intensive* processing. Here the data processing speed limits the performance, and a relatively small amount of on-chip cache memory will provide the required data bandwidth.

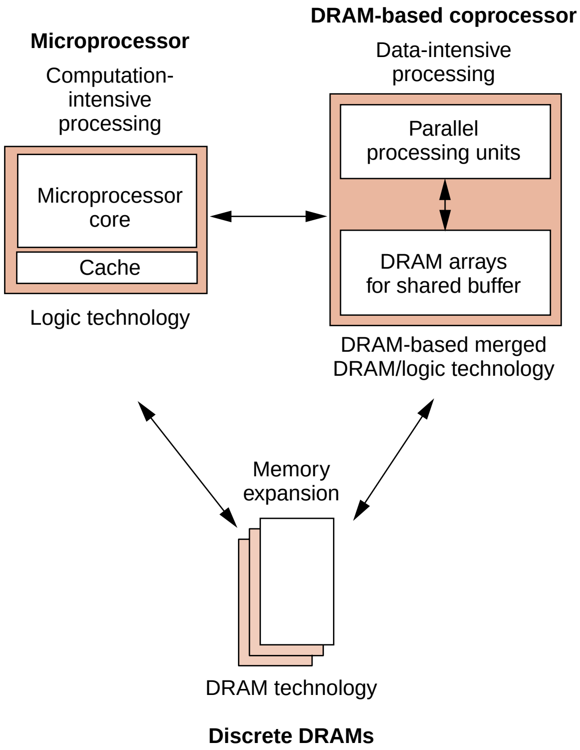
Microprocessor
Computation-intensive processing
Microprocessor core
Cache
Logic technology
DRAM-based coprocessor
Data-intensive processing
Parallel processing units
DRAM arrays for shared buffer
DRAM-based merged DRAM/logic technology
Memory expansion
DRAM technology
Discrete DRAMs

Figure 5. Concept of a DRAM-based coprocessor that works with the microprocessor and (optional) discrete DRAMs for an improved power/performance trade-off.

The DRAM-based coprocessor will contain sufficient memory for the buffers its operation requires and for the microprocessor as well. In this way, the two chips, together with additional DRAM chips, will work in harmony to realize better overall performance. To reduce the overhead between the microprocessor and coprocessor, the design lets the coprocessor have a small microprocessor core and work in a multiprocessor configuration.

Of course, the dream of LSI architects and designers is the emergence of a competitive merged DRAM/logic technology that doesn't compromise transistor performance and DRAM density. The technology is mainly for low-end mobile applications and will be the key to realizing single-chip computers that work for a long time on a small amount of power from a battery, solar cell, or power generator. The rising level of integration is creating a closer bond between computer architecture and semiconductor technology, making it hard to discuss one without considering the other.

THE SIMPLE CELL STRUCTURE, ohmic operating scheme, and continuing progress of CMOS technology suggest that DRAMs will survive for the next 10 years. Beyond that, the major problem for DRAMs will stem from their use of the difference in the Fermi potential across the cell capacitor to dynamically store bit information. The smaller capacitance and larger leakage resulting from further miniaturization will start making it difficult to hold the information. Adoption of material with a higher dielectric constant will help, but eventually some kind of built-in positive feedback mechanism may be needed. The required positive feedback may be either ferroelectric,[22] magnetic, or quantum mechanical.

References

1. J.P. Eckert Jr., "A Survey of Digital Computer Memory Systems," *Proc. IEEE*, Vol. 85, No. 1, Jan. 1997, pp. 184-197. (Originally in *Proc. Inst. Radio Engineers*, Vol. 41, Oct. 1953, pp. 1393).
2. E.W. Pugh et al., "Solid State Memory Development in IBM," *IBM J. Research and Development*, Vol. 25, No. 5, Sept. 1981, pp. 585-602.
3. R. Dennard, Field-Effect Transistor Memory, US Patent 3387286, 1968.
4. R. Dennard, "Evolution of the MOSFET Dynamic RAM—A Personal View," *IEEE Electron Devices*, No. 11, Nov. 1984, pp. 364-369.
5. C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, Reading, Mass, 1979.
6. N.C.-C. Lu et al., "A 22-ns 1-Mbit CMOS High Speed DRAM with Address Multiplexing," *IEEE J. Solid-State Circuits*, Vol. 24, Oct. 1989, pp. 1198-1205.
7. Y. Takai et al., "A 23-ns 1-Mb BiCMOS DRAM," *IEEE J. Solid-State Circuits*, Vol. 25, Oct. 1990, pp. 1102-1111.
8. D. Burskey, "Combination DRAM-SRAM Removes Secondary Cache," *Electronic Design*, Jan. 23, 1992, pp. 39-43; http://www.edram.com/.
9. K. Dosaka et al., "A 100-MHz 4-Mb Cache DRAM with Fast Copyback Scheme," *IEEE J. Solid-State Circuits*, Nov. 1992, pp. 1534-1539.
10. MoSys Inc., http://www.mosys.com/.
11. N. Kushiyama et al., "500 Mbyte/s Data-Rate 512 Kbits × 9 DRAM Using Novel I/O Interface," *Dig. Tech. Papers, 1992 Symp. VLSI Circuit*, June 1992, pp. 66-67; http://www.rambus.com/.
12. D. Burskey, "Dual-Port DRAM Accelerates Windows," *Electronic Design*, Nov. 1, 1993, pp. 43-48; http://www.sec.samsung.com/Products/dram/.
13. K. Inoue et al., "A 10Mb 3D Frame Buffer Memory with Z-Compare and α-Blend Units," *Dig. Tech. Papers, 1995 IEEE Int'l Solid-State Circuits Conf.*, IEEE, Piscataway, N.J., 1995, pp. 302-303.
14. H.S. Stone, "A Logic-in-Memory Computer," *IEEE Trans. Computers*, Jan. 1970, pp. 73-78.
15. K. Sawada et al., "A 72K CMOS Channelless Gate Array with Embedded 1Mbit Dynamic RAM," *Proc. IEEE Custom Integrated Circuits Conf.*, May 1988, 20.3.1-4.
16. T. Shimizu et al., "A Multimedia 32b RISC Microprocessor with 16Mb DRAM," *Dig. Tech. Papers, 1996 IEEE Int'l Solid-State Circuits Conf.*, IEEE, 1996, pp. 262-263.
17. D. Patterson et al., "Intelligent RAM (IRAM): Chips That Remember and Compute," *Dig. Tech. Papers*, *1997 IEEE Int'l Solid-State Circuits Conf.*, IEEE, 1997, pp. 224-225.
18. D. Patterson et al., "A Case for Intelligent RAM," *IEEE Micro*, Mar./Apr. 1997, pp. 34-44.
19. K. Murakami et al., "Parallel Processing RAM Chip with 256Mb DRAM and Quad Processors," *Dig. Tech. Papers, 1997 IEEE Int'l Solid-State Circuits Conf.*, IEEE, 1997, pp. 228-229.
20. S. Przybylski et al., "SDRAMs Ready to Enter PC Mainstream," *Microprocessor Report*, Vol. 10, No. 6, May 6, 1996.
21. P. Gillingham and B. Vogley, "SLDRAM: High-Performance, Open-Standard Memory," *IEEE Micro*, this issue; http://www.SLDRAM.com/.
22. Y. Tarui, "Future DRAM Development and Prospects for Ferroelectric Memories," *Dig. Tech. Papers, 1994 Int'l Electron Device Meeting*, IEEE, 1994, pp. 1.2.1-1.2.10.

Yasunao Katayama is an advisory researcher at IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd. He has been engaged in research on positron physics, quantum devices, high-speed memory design, merged DRAM/logic, and memory architecture.

Katayama received BS and MS degrees in physics from Tokyo University and MS and PhD degrees in electrical engineering from Princeton University. He is a member of the American Physics Society, the IEEE, the Information Processing Society of Japan, and the Institute of Electronics, and Information and Communication Engineers of Japan.

Direct questions or comments about this article to Yasunao Katayama, IBM Research, Tokyo Research Laboratory, 1623-14 Shimotsuruma, Yamato, Kanagawa 242, Japan; katayama@trl.ibm.co.jp.

Reader Interest Survey

Indicate your interest in this article by circling the appropriate number on the Reader Service Card.

Low 153 Medium 154 High 155